BLOQUE 3 – TEMA 9
MODELOS ESTADÍSTICOS Y PROBABILIDAD. LA CURVA NORMAL

Los modelos nos van a permitir comprender la realidad, acercarnos a su explicación, e incluso, tomar decisiones en el campo de la prueba de hipótesis.

Un modelo es un esquema teórico, generalmente en forma matemática de un sistema o una realidad compleja, que se elabora para facilitar su comprensión y el estudio de su comportamiento.

Los modelos pueden surgir de las teorías, pero también puede darse el caso de modelos elaborados a partir de datos observados. Sea un caso ó el otro, el modelo basado en la teoría debe funcionar bien en la realidad, y el modelo surgido de la realidad debe encontrar una teoría que lo respalde.

Entre realidad y modelo debe existir una íntima relación y no cabe pensar en una sin tener en cuenta el otro. Un modelo que no explique con razonable adecuación la realidad, no nos sirve. Una realidad para la que no somos capaces de encontrar un modelo que la represente con fidelidad, nos hace muy difíciles tanto su comprensión como la toma de decisiones en torno a la misma.

El modelo, finalmente, **ha de ser útil**, derivando de una propiedad de la teoría, que es la *comprobabilidad*. Si la teoría es correcta, se podrá contrastar con la realidad.

RPRESENTACIÓN DE LOS MODELOS

El modelo es la representación simplificada de la realidad. A su vez, el modelo puede representarse de varias formas:

- ◊ representación **icónica** → una escultura, un cuadro, etc.
- ◊ representación **matemática** → mediante fórmulas se establece una igualdad más o menos compleja.
- ◊ representación **analógica** → con esquemas, diagramas, etc., podemos entender la realidad abstracta y muchas veces solo observable mediante el uso de aparatos complejos. Por ejem.: el sistema solar, el átomo.

Efectuar predicciones de futuro acerca de la conducta humana es complicado. Si somos capaces de establecer modelos suficientemente cercanos a la realidad, podremos hacer este tipo de predicciones, tales como el porcentaje de suspensos, o de niños violentos, o de hiperactivos, etc.

Modelos matemáticos y modelos estadísticos

Un modelo estadístico es un tipo de modelo matemático en el que se incorpora como componente fundamental la probabilidad.

Los modelos matemáticos se expresan mediante igualdades que reflejan la relación existente entre los componentes de la realidad. La función es un modelo matemático. Un **modelo matemático clave** en nuestro campo es la **campana de Gauss**, también conocida como **distribución normal tipificada**.

$$P(x) = \frac{1}{\sqrt{2\Pi\sigma}} . e^{\frac{-1}{2} . \left(\frac{\mu - \bar{x}}{\sigma}\right)^2} = \frac{1}{\sqrt{2\Pi\sigma}} . e^{\frac{-1}{2} . z^2}$$

Los valores de cualquier ordenada pueden calcularse para cada distribución empírica de N datos, conociendo el valor de su $\bar{x}$ y de la μ y la σ de la población.

Toda predicción asume un cierto riesgo de error, error que consideraremos aleatorio. Los errores aleatorios tienden a compensarse y su magnitud puede estimarse. Como podemos imaginar, nuestro interés consiste en reducir a su mínima expresión este error, siendo la meta a alcanzar que $e = 0$. La expresión, ya vista, $X = V \pm E$, se conoce como Modelo Lineal Clásico.

PROBABILIDAD

El principio básico de las pruebas estadísticas de significación consiste en **comparar los resultados obtenidos con los esperados por azar**. Cuando los resultados de una prueba estadística superan a lo esperado por puro azar, se aceptará que el fenómeno en cuestión no se explica por el azar sino por la acción del investigador, sometida a contraste en condiciones controladas y analizada mediante tal prueba.

Diferenciamos entre **dos conceptos de probabilidad**:

- probabilidad ***a priori*** $\rightarrow$ antes de que ocurra un fenómeno podemos estimar las probabilidades de que ocurra ó de acertar una predicción.
- probabilidad ***a posteriori*** $\rightarrow$ cuando los fenómenos ya han ocurrido, podemos establecer la probabilidad de ocurrencia de tal fenómeno. Esta probabilidad se calcula empíricamente y se traduce en la frecuencia relativa con la que ocurre tal fenómeno cuando se repite un elevado número de veces en las mismas condiciones.

Probabilidad a priori y a posteriori

La probabilidad a priori se establece sobre la base del nº de casos favorables dividido por el nº de casos posibles, y constituye la base de la probabilidad matemática, de la probabilidad en términos teóricos.

En el caso de **la probabilidad a posteriori**, cuando estudios reiterados vienen a arrojar resultados compatibles, es posible establecer la realización de estimaciones sobre la probabilidad de que, en ocasiones sucesivas, ocurra un determinado acontecimiento.

Cálculo de la probabilidad. Definiciones

ESPACIO MUESTRAL $\rightarrow$ es el conjunto de todos los resultados posibles de un fenómeno. Cuando dos conjuntos, A y B, no tienen elementos en común, decimos que $A \cap B = \varnothing$; y se lee "A intersección con B, es igual al conjunto vacío". Entonces:

$$P(A \cup B) = P(A) + P(B)$$

Pero esto no es lo normal, pues se da con bastante frecuencia que dos ó más conjuntos tengan elementos en común y entonces decimos que $A \cap B \neq \varnothing$, luego

$$P(A \cup B) = P(A) + P(B) - P(A \cap B)$$

El fenómeno de la ***EXHAUSTIVIDAD*** ó ***AGOTAMIENTO*** se produce cuando los diferentes subconjuntos que puedan crearse, son subconjuntos del espacio muestral y todos ellos juntos lo agotan, lo completan plenamente.

Hablamos de ***MUTUA EXCLUSIÓN*** cuando dos acontecimientos distintos no tienen ningún elemento en común, es decir, su intersección es el $\varnothing$. Como en tal caso $A \cap B = \varnothing$, las probabilidades de cada subconjunto pueden sumarse.

LA INDEPENDENCIA supone que la probabilidad de que ocurra el fenómeno conjunto, es igual al producto de las probabilidades de cada uno por separado. Entonces:

$$P(A \cap B) = P(A) \cdot P(B)$$

Esta es una de las condiciones o supuestos que se han de verificar para la aplicación de las pruebas denominadas paramétricas en el contraste de hipótesis.

Concepto de ***PROBABILIDAD CONDICIONAL***.

En nuestro ámbito de trabajo es bastante frecuente encontrarnos con fenómenos relacionados, es decir, que no son independientes. Como en este caso, $A \cap B \neq \varnothing$, la probabilidad condicional nos sitúa ante un caso en el que deseamos conocer la probabilidad de un determinado acontecimiento ó suceso cuando la probabilidad del otro es conocida.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Variables aleatorias continuas y discretas

- Una variable cuantitativa es **discreta** cuando no puede adquirir todos los valores posibles, es decir, cuando el conjunto de sus valores es numerable.
- Una variable cuantitativa es **continua** cuando admite un número "no numerable" de casos o valores.
- ver ejem. en pág. 186.

Si realizamos gráficos de histogramas, disminuyendo progresivamente la base de cada barra, como se aprecia en la fig. 9.1, pág. 187, llegaríamos a una curva que uniría los infinitos puntos posibles de esa distribución. Esas sucesivas representaciones gráficas nos permiten desvelar dos valores de la probabilidad: la **función de densidad de probabilidad**, y la **función de distribución**. Tienen estas características:

- el área ocupada por ambas representaciones tiene como valor la unidad.
- en todos los casos, las representaciones tienen siempre valores positivos.
- en las representaciones de histogramas de barras, y sea cual sea su base, el eje de abscisas representa una variable aleatoria discreta.
- en las representaciones de histogramas sin barras, cuando la base son puntos y la representación es una curva, estamos ante una variable aleatoria continua.

En el caso de las variables aleatorias discretas, el área que queda entre dos valores A y b, nos indica la PROPORCIÓN de casos del total que se encuentran entre esos valores. En el caso de variables cuantitativas continuas, indica la PROBABILIDAD de que tal variable tome esos valores.

ALGUNAS FUNCIONES DE DENSIDAD DEPROBABILIDAD

Función de densidad de probabilidad normal

La más frecuentemente utilizada por nosotros es la denominada normal, conocida como campana de Gauss. Ver fig. 9.2 pág. 189 para una representación de la curva. Algunas de sus características más importantes:

- porcentaje de casos entre dos valores de σ (desviación típica poblacional):
  - $\pm\sigma \rightarrow 68\%$ $\pm2\sigma \rightarrow 95\%$ $\pm4{,}5\sigma \rightarrow 99{,}99932\%$
  - estos valores en porcentajes se convierten en *probabilidades* dividiendo por 100; es decir, el conjunto de los casos situados entre $\pm\sigma$ tiene una probabilidad de ocurrencia de 0,68
- mediante la **tabla de áreas de la curva normal**, podemos atribuir probabilidades a un caso concreto, en sus diversas columnas.

Función de densidad de probabilidad χ^2

Las variables aleatorias a las que se les aplica se distribuyen según χ^2 con **n-1** grados de libertad (υ), esto es, el nº de casos menos 1, puesto que mientras los primeros casos pueden variar libremente, el último viene condicionado por todos los anteriores.

Según aumenta el nº de los grados de libertad, la distribución χ^2 se aproxima progresivamente a la distribución normal (ver fig. 9.3 pág. 190). De hecho, las tablas de χ^2 nos ofrecen valores de probabilidad hasta 30 gl. A partir de ahí, la distribución sigue con un valor de:

$$z = \sqrt{2\chi^2} - \sqrt{2(gl) - 1}$$

con $\bar{x} = 0$ y $S = 1$

La importancia de esta distribución radica en sus aplicaciones:

- ◊ como prueba de bondad de ajuste
- ◊ como prueba de independencia
- ◊ como prueba del grado de asociación entre dos conjuntos de variables de atributo, para calcular el coeficiente de contingencia C:

$$C = \sqrt{\frac{\chi^2}{N + \chi^2}}$$

Función de densidad de probabilidad t

La distribución de probabilidad t se conoce como t de Student. Si tenemos Y y Z, dos variables aleatorias independientes, Y con una distribución χ^2 con **n** gl., y Z con una distribución normal (o,1), definimos la distribución

$$t = \frac{Z}{\sqrt{\frac{Y}{\upsilon}}}$$

Cuando aumentan los gl., la distribución se aproxima progresivamente a la campana de Gauss (ver fig. 9.4 pág. 191).

Esta distribución **se utiliza frecuentemente en pruebas de contraste de hipótesis** para decidir si la diferencia de medias es o no es estadísticamente significativa, a un determinado nivel de confianza.

Cuando las muestras son correlacionadas, la distribución t no sigue el estadístico de contraste aplicado en las muestras independientes. Dos muestras son correlacionadas cuando se forman parejas de sujetos, uno de cada muestra, que gozan de cierta característica común o similar.

Función de densidad de probabilidad F

Denominada F de Fisher, esta función de probabilidad, junto con la anterior t, son de las más utilizadas en el ámbito del contraste de hipótesis.

$F = t^2$

F puede aplicarse además a contrastes con tres o más pares de medias en diseños de tres o más grupos. **Se aplica fundamentalmente en el análisis de varianza (ANAVA).**

F nos indica si se dan o no diferencias estadísticamente significativas entre varios grupos de medias. En caso afirmativo, es preciso averiguar entre qué dos pares de medias se concreta tal diferencia, por lo que se hace necesario la continuación del trabajo con las denominadas **pruebas a posteriori**.

La distribución F se define como la razón entre dos distribuciones χ^2 independientes, dividida cada una de ellas entre sus respectivos gl.

$$F = \frac{\chi_1^2/\upsilon_1}{\chi_2^2/\upsilon_2}$$

Si las dos varianzas poblacionales son iguales, la fórmula se reduce a:

$$F = \frac{s_1^2}{s_2^2}$$

Las tablas nos ofrecen los valores de probabilidad que corresponden a diversos gl del numerador (varianza **INTERgrupos en la ANAVA**) y del denominador (varianza **INTRAgrupos en la ANAVA**). La distribución F es no negativa, sesgada hacia la derecha y sus valores oscilan entre 0 e ∞, siendo asintótica al eje de abscisas.

LA CURVA NORMAL DE PROBABILIDADES

En ciencia no es aconsejable hacer apreciaciones subjetivas acerca del grado de acercamiento o parecido entre el modelo y la realidad. Conviene que las apreciaciones sean precisas y para ello tenemos la **prueba de bondad de ajuste**.

Si los datos empíricos se ajustan razonablemente al modelo, es decir, si las discrepancias son compatibles con las esperadas por puro azar, consideraremos estos datos como normales, y les aplicaremos todas las propiedades del modelo. En caso contrario, no podemos afirmar que el modelo sea idóneo.

Sobre el modelo

No hay una única curva normal, sino una por cada par de valores de $\bar{x}$ y de S. Todas ellas cumplen las siguientes **características** (ver fig. 9.2 pág. 198):

- el valor máximo de la ordenada corresponde a la $\bar{x}$, y por tanto, a una $z = 0$
- a ambos lados de la $\bar{x}$, se encuentran dos puntos de inflexión que se corresponden con los valores de $z \pm 1$
- la curva es simétrica respecto a la $\bar{x}$, puesto que $\bar{x}$ = Md = Mo. La ordenada de la $\bar{x}$ divide a la curva en dos partes iguales, cada una con el 50% de los casos
- la curva es asintótica al eje de abscisas, por lo que nunca se llega a abarcar el 100% de los casos.

La prueba de bondad de ajuste

Pretendemos poder afirmar que una distribución dada se desarrolla normalmente, o que sus datos se distribuyen normalmente, siguiendo el modelo de la normal (0,1).

Para ello se acude a las pruebas de bondad de ajuste. Nosotros utilizaremos la de χ^2

Esta prueba valora las discrepancias entre los valores de las frecuencias empíricas y las teóricas. Si estas discrepancias no fueran significativas a un determinado nivel de confianza, admitiríamos que los datos empíricos y el modelo son una misma cosa (aceptamos Ho). Admitiríamos que las discrepancias encontradas pueden explicarse por puro azar como consecuencia de errores de muestreo.

$$\chi^2 = \sum \left[\frac{(f_0 - f_e)^2}{f_e}\right]$$

Se distribuye según χ^2 con:

* gl = nº de columnas -1 (si μ y σ son conocidas)
* gl = nº de columnas -3 (si μ y σ son estimadas)

Ver tabla 9.2 pág. 197

Xi	**fo**	**Lsup**	**Zi**	**P(Zi)**	**Pi**	**fe**	**(fo-fe)**	**(fo-fe)²**	$\frac{(fo-fe)^2}{fe}$
	$\sum$					$\sum$			$\sum$

Xi → puntuaciones directas obtenidas por el alumno
fo → frecuencias observadas
Lsup → límite superior de cada intervalo
Zi → las puntuaciones típicas de tales límites.
(Necesitamos calcular previamente los valores de $\bar{x}$ y de S)
P(Zi) → probabilidades que corresponden a esas Zi (columna B en las tablas)
Pi → probabilidad de cada intervalo, que se calcula restando de su valor P(Zi), el valor que corresponde al intervalo anterior.
fe → frecuencias esperadas o teóricas. Para su cálculo:
* hallamos la columna P(Zi)
* hallamos la columna Pi
* hallamos la columna fe → Pi . N

Calculado el valor de χ^2 empíricamente, lo comprobamos con el valor de las tablas, para un nivel de confianza y (n-1) gl.

Si $\chi^2 \leq$ valor de tablas → acepto Ho → no hay diferencias significativas entre las distribuciones → el ajuste es bueno
Si $\chi^2 >$ valor de tablas → acepto H1 → si hay diferencias significativas entre las distribuciones → no hay ajuste.